

<b>REPORT DOCUMENTATION PAGE</b>				Form Approved OMB No. 0704-0188
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p><b>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</b></p>				
1. REPORT DATE (DD-MM-YYYY) 17-03-2015		2. REPORT TYPE Final		3. DATES COVERED (From - To) 06-03-2012 – 07-03-2015
<b>4. TITLE AND SUBTITLE</b>  Model Learning for Probabilistic Simulation on Rare Events and Scenarios			5a. CONTRACT NUMBER FA2386-12-1-4053	
			5b. GRANT NUMBER Grant AOARD-124053	
			5c. PROGRAM ELEMENT NUMBER 61102F	
<b>6. AUTHOR(S)</b>  Takashi Washio			5d. PROJECT NUMBER	
			5e. TASK NUMBER	
			5f. WORK UNIT NUMBER	
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b>  The Institute of Scientific and Industrial Research, Osaka University 8-1, Mihogaoka, Ibarakiishi, Osaka, 5670047, Japan			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>  N/A	
<b>9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>  AOARD UNIT 45002 APO AP 96338-5002			<b>10. SPONSOR/MONITOR'S ACRONYM(S)</b>  AFRL/AFOSR/IOA(AOARD)	
			<b>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</b>  AOARD-124053	
<b>12. DISTRIBUTION/AVAILABILITY STATEMENT</b>  Distribution Code A: Approved for public release, distribution is unlimited.				
<b>13. SUPPLEMENTARY NOTES</b>				
<b>14. ABSTRACT</b> This project established a new methodology for probabilistic inference and prediction of rare and special events and/or scenarios based on the simulation models and the observed data that rarely contain rare events, applied it to a rainfall flood risk analysis of Chikugo river, Japan, and showed that it can generate various rainfall scenario that causes a flood. Rainfall pattern that causes a flood is generated by Replica Exchange Monte Carlo algorithm, and covariant shift phenomenon was corrected by placing more weight on the flood region. This work gives a general framework to cope with the problem of handling a complex and/or large scale system where a complete set of possible events and scenarios is hardly obtained.				
<b>15. SUBJECT TERMS</b> Rare event, Data mining, Probabilistic simulation, Sample generation, Replica Exchange Monte Carlo, Covariance shift, Flood risk analysis, Metropolis algorithm, Markov Chain Monte Carlo				
16. SECURITY CLASSIFICATION OF: a. REPORT    b. ABSTRACT    c. THIS PAGE		17. LIMITATION OF ABSTRACT SAR	18. NUMBER OF PAGES 15	19a. NAME OF RESPONSIBLE PERSON Hiroshi Motoda, Ph. D.
U                  U                  U				19b. TELEPHONE NUMBER (Include area code) +81-42-511-2011

Report Documentation Page			Form Approved OMB No. 0704-0188	
<p>Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p>				
1. REPORT DATE <b>26 MAR 2015</b>	2. REPORT TYPE <b>Final</b>	3. DATES COVERED <b>06-03-2012 to 07-03-2015</b>		
4. TITLE AND SUBTITLE <b>Model Learning for Probabilistic Simulation on Rare Events and</b>			5a. CONTRACT NUMBER <b>FA2386-12-1-4053</b>	
			5b. GRANT NUMBER	
			5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S) <b>Takashi Washio</b>			5d. PROJECT NUMBER	
			5e. TASK NUMBER	
			5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) <b>The Institute of Scientific and Industrial Research,,Osaka University,8-1, Mihogaoka,Ibarakishi,,Osaka,NA,NA</b>			8. PERFORMING ORGANIZATION REPORT NUMBER <b>N/A</b>	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) <b>AOARD, UNIT 45002, APO, AP, 96338-5002</b>			10. SPONSOR/MONITOR'S ACRONYM(S) <b>AFRL/AFOSR/IOA(AOARD)</b>	
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) <b>AOARD-124053</b>	
12. DISTRIBUTION/AVAILABILITY STATEMENT <b>Approved for public release; distribution unlimited</b>				
13. SUPPLEMENTARY NOTES				
14. ABSTRACT <p><b>This project established a new methodology for probabilistic inference and prediction of rare and special events and/or scenarios based on the simulation models and the observed data that rarely contain rare events, applied it to a rainfall flood risk analysis of Chikugo river, Japan, and showed that it can generate various rainfall scenario that causes a flood. Rainfall pattern that causes a flood is generated by Replica Exchange Monte Carlo algorithm, and covariant shift phenomenon was corrected by placing more weight on the flood region. This work gives a general framework to cope with the problem of handling a complex and/or large scale system where a complete set of possible events and scenarios is hardly obtained.</b></p>				
15. SUBJECT TERMS <b>Rare event, Data mining, Probabilistic simulation, Sample generation, Replica Exchange Monte Carlo, Covariance shift, Flood risk analysis, Metropolis algorithm, Markov Chain Monte Carlo</b>				
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT <b>Same as Report (SAR)</b>	18. NUMBER OF PAGES <b>16</b>
a. REPORT <b>unclassified</b>	b. ABSTRACT <b>unclassified</b>	c. THIS PAGE <b>unclassified</b>	19a. NAME OF RESPONSIBLE PERSON	

**“Model Learning for Probabilistic Simulation on Rare Events and Scenarios”**

**March, 6th, 2015**

**Name of Principal Investigators (PI and Co-PIs):**

- e-mail address : washio@ar.sanken.osaka-u.ac.jp
- Institution : The Institute of Scientific and Industrial Research, Osaka University
- Mailing Address : 8-1, Mihogaoka, Ibaraki, Osaka, 5670047, Japan
- Phone : +81-6-6879-8540
- Fax : +81-6-6879-8544

Period of Performance: March/7/2012 - March/6/2015

**Abstract:** Studies of probabilistic model learning reflecting data and background knowledge have been performed extensively. Nevertheless, few reports have described conditional probabilistic model learning methods under rare and special conditions. This study aims to establish a new methodology for probabilistic inference and prediction of rare and special events and/or scenarios based on their simulation models, and further aims to develop a principle to learn the simulation models efficiently and accurately from known data and background knowledge.

In the first year of this research project, we established a highly efficient methodology to simulate a target rare and special event/scenario and quantitatively evaluate its probability by using its conditional probability model. Given a simulation model of our target rare and special event/scenario based on our background knowledge, we synthesized it with probability distribution of our interesting condition having relevance with the target event/scenario empirically observed in the past. This synthesis quantitatively provides the probability distribution of our interesting condition which causes our target rare and special event/scenario through replica exchange Monte Carlo (REM) algorithm. We applied this new methodology to estimate rare rain fall scenarios and their probability distribution which cause severe floods of Chikugo River in Kyushu Island in Japan. Our advanced scheme successfully derived the distribution of the dangerous rain fall scenarios and their probability under simulation based experiments together with empirical rain fall data observed in the last 10 years.

In the second year, we extended this approach to enhance the credibility of our simulation and evaluation. Because the data we ordinary observe in the past hardly include the records causing the target rare event/scenario, its probability tends to be underestimated, if we use the past observed data only. This is a problem called covariate shift in statistics. We need to calibrate the probability to avoid the underestimation of our target risk induced by this problem. We overcome this issue by calibrating the rain fall distribution based on our simulation model which reflects our prior knowledge on the target event/scenario. We applied this extended approach to the aforementioned problem of Chikugo River, and evaluated its effect by comparing the results with these obtained in the first year. The comparison indicated the importance of this calibration, since significant increases of the probability of the target event/scenario were found in some cases.

In the last year, we integrated the approaches of the simulation, the evaluation, the synthesis and the calibration into a unified computation scheme, and applied it to analyze more complicated and detailed events/scenarios of the severe floods of Chikugo River. We analyzed the spatial and temporal distributions of the severe rain falls in the basins of Chikugo River and its tributaries, and showed possible events/scenarios of the rain fall distributions causing the floods in this area. Moreover, we demonstrated the intervals of the river where the infrastructures for the flood control are insufficient. The outcomes provided in this study have significant contributions to risk assessment of natural and/or artificial disasters. Not only to this application, the technique developed in this project is considered to have a big impact to various domains of science and engineering, since researches and assessments in such domains often require extrapolative estimations and predictions of special conditions in their interests.

## **Introduction:**

Needs for learning of probabilistic simulation model of rare and special event/scenario occurrences have been increasing. Meeting those needs is expected to cope with chemical reactions that are highly specific, but proceed with high probability through numerous thermal trial runs, as represented by formation of DNA double helix or protein folding structures, and to cope with prediction of damage by massive earthquakes, giant tsunami and sever river floods which are experienced only rarely or never throughout the history of humankind.

However, known data and background knowledge at our hands do not sufficiently cover information required for developing the accurate and reliable simulation model of such rare and special events and scenarios. Probabilistic distributions of events and scenario occurrence under rare and special conditions are, in most cases, vastly different from those that exist under ordinary conditions. For this reason, voluminous simulations using probabilistic models learned from such known data and background knowledge are unable to estimate statistically sufficient distributions related to events and scenarios meeting the target conditions. Under this limitation, the current model based simulation techniques hardly provide any realistic estimates and predictions in practical applications where some rare and special event and scenario occurrences provide their main consequences. Therefore, we intend to develop the followings in this project for three years:

- (1) Highly-efficient simulation principles of conditional probabilistic distributions targeted by that model.
- (2) Highly-accurate learning principles of probabilistic simulation models under target conditions inferred from known data and background knowledge.
- (3) Methods of integrated realization of these principles.

Furthermore, we intend to accomplish the following:

- (4) Adaptation of the developed methods to actual applications such as simulations of rare mega-disaster scenarios.
- (5) Practical demonstrations of the adapted methods and establishments of new simulation methodologies in the practical domains such as risk assessment.

In the first year, we particularly concentrated our research effort to the objective (1) which outcome provides a basis of the other objectives. In addition, we evaluated our developed method for the objective (1) through its application to risk assessment of severe river flood which is a representative but rare natural disaster in progressive countries where infrastructures for the river flood control are well developed. This performance evaluation of our developed method addresses some part of our objectives (4) and (5).

In the second year, we worked on the objective (2) which is to extend the outcome in the first year for enhancing the credibility of the simulation model. The probability of the severe rain fall causing the target rare events/scenarios tends to be underestimated, if we use the past observed data only. This is because the rain fall data observed in a past limited period hardly include the severe rain fall records causing the events/scenarios. This is called a covariate shift problem in statistics. We need to calibrate the probability to avoid the underestimation induced by this problem. We also applied the extended approach to the risk assessment of the severe river floods similar to the first year's. Its performance has been compared with the results with that obtained in the first year. This evaluation task belongs to some part of our objectives (4) and (5).

In the last year, we integrated the aforementioned approaches of the simulation, the evaluation, the synthesis and the calibration into a unified computation scheme for the probabilistic simulation of rare events/scenarios. This is to attain the objective (3). We farther applied the integrated scheme to analyze more complicated and detailed events/scenarios of the severe floods of Chikugo River. This addresses our objectives (4) and (5) and enabled the detailed analyses on the spatial and temporal distributions of the severe rain falls in the basins of Chikugo River and its tributaries. The results demonstrated the possible events/scenarios of the rain fall distributions causing the floods in this area and clearly suggested the intervals of the river where the infrastructures for the flood control are insufficient.

## Principles and Methods:

### (1) Background and Basic Simulation Principles

Studies of rare event simulation using given probabilistic models such as cross-entropy method [1], multi-canonical method [2], and replica exchange method [3] have widely been undertaken within conventional Monte Carlo simulation approaches, and they have been used in varieties of areas [4, 5, 6, 7]. However, the method of efficient and accurate simulations by combining probabilistic models of target events/scenarios under rare, special conditions and observation data associated with the target events/scenarios has remained almost entirely unexamined.

In the first year, we developed a novel framework for the objective (1) which introduces empirical probability distribution of variables conditioning our target events and scenarios to a Monte Carlo simulation model of the targets. Then, a conditional probability distribution of the conditioning variables under the occurrence of the target events and scenarios are derived by a Bayesian estimation. Mathematically, a probabilistic simulation model  $P(X|S; \Theta_{X|S})$  of random variable vector  $X$ , which consists of the variables conditioning the target rare and special events/scenarios, under a certain conditional variable vector  $S$ , representing the target events and scenarios, and its parameter  $\Theta_{X|S}$  as shown at the upper part in Figure 1 is inferred with maximum likelihood from background knowledge and data. If a joint probability distribution of  $X$  and  $S$  is determined using two probability distributions  $P(X; \Theta_X)$  and  $P(S|X; \Theta_{S|X})$ , as shown in the lower part in Figure 1, where the former is from our empirically observed data and the latter is a simulation model provided from our background knowledge, then  $P(X|S; \Theta_X, \Theta_{X|S})$  is presumed by Bayes' theorem as expressed by the equation shown at the bottom of the figure. This provides the probability distribution of the variables:  $X$  conditioning the target rare and special events and scenarios:  $S$  under their occurrences.

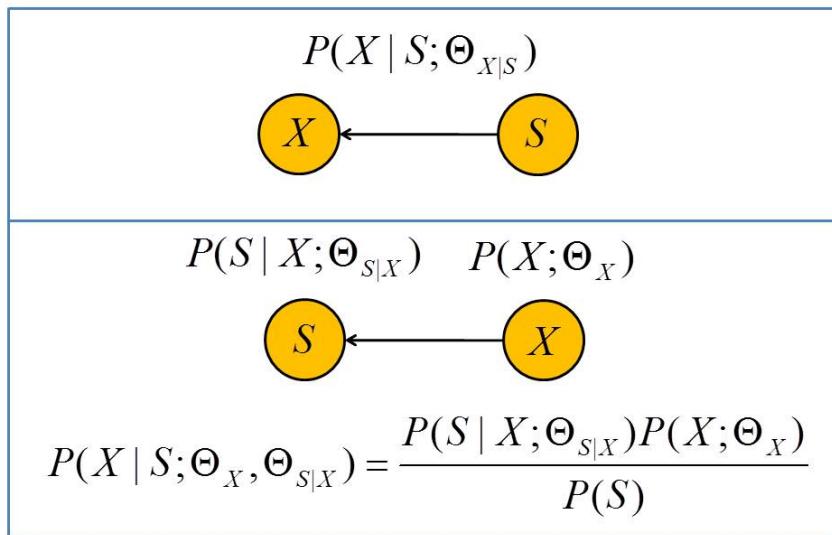


Figure 1 Ordinary probabilistic model and Bayesian probabilistic model.

A main technical issue in this framework is that  $X$  drawn  $P(X; \Theta_X)$  has nearly zero  $P(X|S; \Theta_X, \Theta_{X|S})$  under a rare and special condition  $S$ , and numerous  $X$  generated by or observed from  $P(X; \Theta_X)$  will be wasted during the simulation. Therefore, a principle of the probabilistic inference, which does not distort the distribution  $P(X; \Theta_X)$  and yet generates and uses  $X$  having significantly large  $P(X|S; \Theta_X, \Theta_{X|S})$  efficiently, has to be introduced. As an efficient measure to overcome this issue, we applied the replica exchange method (REM) [3].

The REM consists of two basic principles. One is the Markov Chain Monte Carlo (MCMC) principle [8] for probabilistically and repeatedly generating many conditioning vectors  $X$  following the probability distribution  $P(X|T_k, S; \Theta_X, \Theta_{X|S})$  which is usually complex and prohibits any analytical generation of  $X$ . Here,  $T_k$  is an artificially introduced parameter named "temperature" and used in the second principle to control the rareness and specialty of the target events and scenarios. The MCMC principle applies a generate and test algorithm named Metropolis algorithm [9] to maintain the

distribution of the generated  $X$  to be  $P(X|T_k, S; \Theta_X, \Theta_{X|S})$ .

The second is the replica exchange principle. We generate the conditioning vectors  $X$  in parallel MCMC computations under some different temperatures  $T_k$ . These parameters are designed to increase the probability of the target rare and special events/scenarios under its larger value. Accordingly, more target events/scenarios are generated in the MCMC computations having higher temperature  $T_k$  except the MCMC having  $T_0$  which provide the original probability distribution  $P(X|T_0, S; \Theta_X, \Theta_{X|S}) = P(X|S; \Theta_X, \Theta_{X|S})$  we are interested in. The replica exchange principle applies occasional exchanges of the generated  $X$  between the two MCMC computations having neighbor temperatures, i.e.,  $T_{k-1}$  and  $T_k$ . These pairs of MCMC computations are randomly chosen, and the exchange is probabilistically accepted by the Metropolis algorithm. This algorithmic scheme maintains the probability distribution of our interesting conditional vectors  $X$  to be  $P(X|S; \Theta_X, \Theta_{X|S})$  for  $T_0$  based on the nature of the Metropolis algorithm while inserting some  $X$  causing the rare and special events and scenarios from the distribution of  $X$  generated under higher temperatures. Accordingly, we efficiently obtain a set of  $X$  causing the target events/scenarios strictly following  $P(X|S; \Theta_X, \Theta_{X|S})$ .

## (2) Extension of Principles and Method

In the second year, we extended the aforementioned approach to enhance the credibility of our simulation and evaluation. A major effect to reduce the credibility is considered to be statistical covariate shift [10]. In the former analysis, we assumed that the parameters  $\Theta_X$  of the probability distribution  $P(X; \Theta_X)$  observed in a past period always agrees with its true parameters. However, it may be different depending on the rain fall amount as exemplified in Fig.2. The dashed line is  $P(X; \Theta_X)$  in a log scale and its extrapolation up to the region of the inexperienced severe rain fall. In reality, the probability distribution may be  $P(X; \Theta'_X)$  having a different parameter vector  $\Theta'_X$  in the severe region as indicated by a red curve, if the parameters has some dependency to the rain fall amount. This type of scale dependency of the probability distribution is frequently observed in natural events. If we simply use  $\Theta_X$  as the analysis in the first year, its error propagates to the evaluation of  $P(X|S; \Theta_X, \Theta_{X|S})$  through the Bayesian estimation indicated in Fig.1. This is called a covariate shift problem in statistics. Thus,  $\Theta_X$  must be appropriately calibrated into  $\Theta'_X$  in the analysis of the severe region to avoid under- or over-estimation of the target risk.

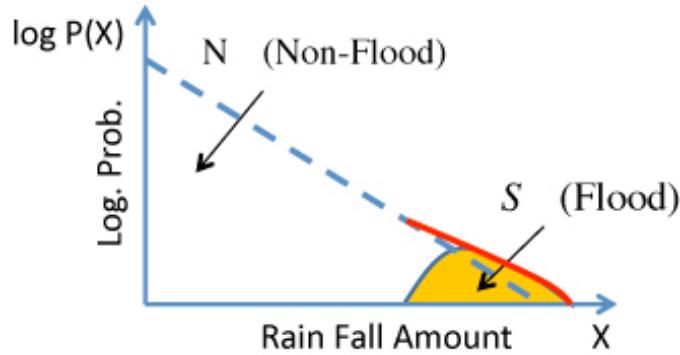


Figure 2 Difference between assumed and real probability distributions.

If we had the rain fall data observed in the severe region, we could estimate  $\Theta'_X$  in a straight manner. But, the main difficulty to address the covariate shift problem is that we can hardly obtain such inexperienced data. To overcome this issue, we proposed a novel extension to use our simulation model,  $P(S|X; \Theta_{S|X})$ , which reflects our prior knowledge on the target event or scenario. Let  $N$  be the condition that our target event or scenario does not occur. By its definition,  $P(S|X; \Theta_{S|X}) + P(N|X; \Theta_{S|X}) = 1$  holds. We obtain the probability distribution  $P(X; \Theta_X) = P(X, N)$  from the observed data, because we almost always do not have the observed data under the rare and special events or scenarios. By the definition of the conditional probability,  $P(N|X) = P(X, N)/P(X)$ , we derive the true  $P(X)$  as

$$P(X) = P(X, N)/P(N|X) = P(X; \Theta_X)/P(N|X; \Theta_{S|X}) = P(X; \Theta_X)/(1 - P(S|X; \Theta_{S|X})).$$

This true  $P(X)$  is  $P(X;\Theta'_X)$  in the severe region and almost  $P(X;\Theta_X)$  otherwise. The parameter vectors  $\Theta_X$  and  $\Theta_{S|X}$  are given by the maximum likelihood estimation over the observed data and our prior knowledge, respectively, as demonstrated in the first year. We replace  $P(X;\Theta_X)$  with this  $P(X)$  in the Bayesian estimation depicted at the bottom of Fig.1 and perform the rare and special events or scenarios similarly to the analyses in the first year.

### (3) Integration of All developed Methods

In the final year, we integrated all methods including the Bayesian inference in Fig.1, its extension to the correction of the covariate shift depicted in Fig.2, the Markov Chain Monte Carlo (MCMC) and the replica exchange method (REM) into a computation scheme. This integration is not only the unification of the computation procedure but adaptation of the computation to stochastically generate the events/scenario under  $P(X;\Theta_X)/(1-P(S|X;\Theta_{S|X}))$  by applying the replica exchange method (REM) to efficiently remove the target event S in the model simulation.

## Experiments and Results:

### (1) Experiments and Results of the Basic Method

In the first year, we applied our developed basic method to estimate extraordinary rain fall scenarios and their probability distribution which cause severe floods of Chikugo River in Kyushu Island in Japan for the objective (4) and (5). This type of risk assessment is highly important for every progressive country. Because infrastructures for the flood control are usually well developed and maintained in these countries, further investment on the infrastructure should be efficiently and scientifically decided based on some objective and quantitative information of the flood risk to save the social investment cost.

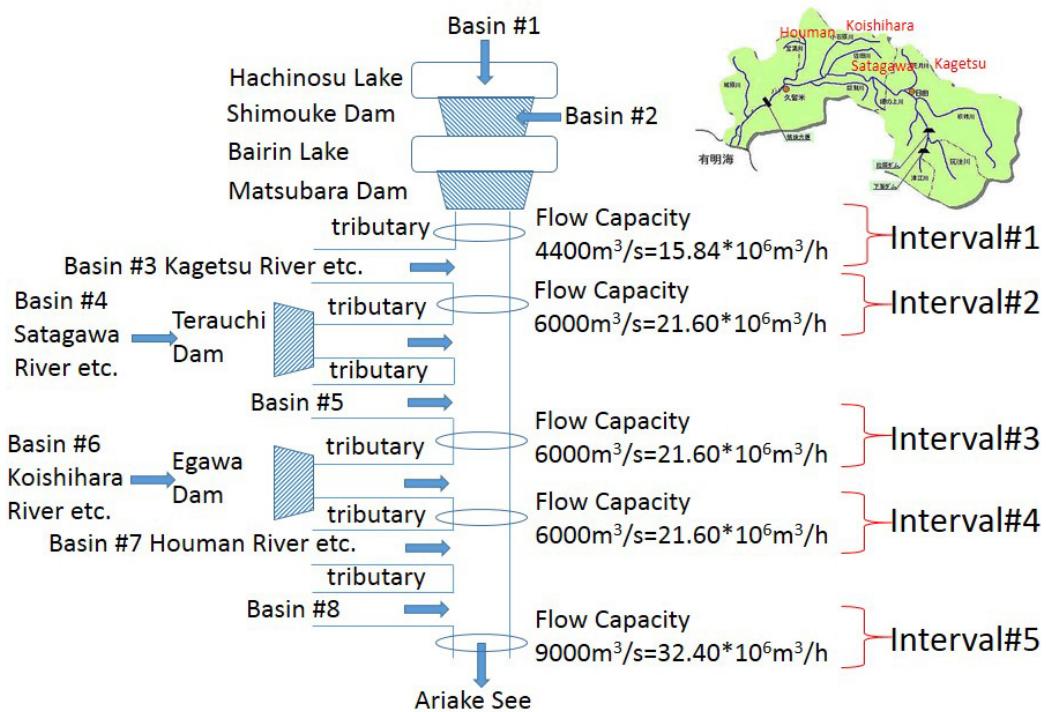


Figure 3 Geological structure of Chikugo river and its maximum flow capacity in every interval.

First, we defined a condition S of a sever flood and its conditioning vector X. Figure 3 depicts the entire geological structure of Chikugo river and its maximum flow capacity of the river. Chikugo river is partitioned into five intervals for controlling the water flow rate by local governments. A maximum flow rate is defined in each interval based on the design specification of the river infrastructure. Accordingly, S is defined that water flow rate per hour exceeds its designed maximum limitation at some place of the river. On the other hand, Chikugo river has four major dams, and daily rain fall

amount is observed and recorded at each dam. In addition, the maximum time period of continuous rain fall was found to be 7 days (a week) by inspecting the rain fall records in the last 10 years. Based on these facts,  $X$  is defined as a time series of daily rain fall amount at the 4 dams for a week.  $X$  is represented by 7 consecutive steps of 4 dimensional vectors. A rain fall scenario represented by  $X$  primarily dominates the river flow and the floods.

Next, we developed a probabilistic simulation model  $P(S|X; \Theta_{S|X})$  which is the probability of occurrence of a severe flood:  $S$  under a given rain fall scenario:  $X$ . Entire basin of the river is partitioned into 8 basins as shown in Fig.3. Each basin is represented by a tank model where a basin is considered to have 3 layered water retention and flows, *i.e.*, these of surface, surface permeation and underground. The 3 layered retention is represented by 3 cascaded tanks, and the flow at each layer is represented by a pipe. Each dam is also represented by a tank, and it is assumed to always take a standard operation mode for heavy rain fall in which its outlet is closed except when it is full. Once it becomes full, its outflow is maintained to balance with its inflow to avoid the overflow. Water flow rate of each river stream is computed based on mass balance of the water flow.

Finally, we set up the probability distribution  $P(X; \Theta_X)$ . All rain fall data recorded at the 4 dams in the last 10 years are preprocessed in form of the 7 consecutive steps of 4 dimensional vectors to represent the rain fall scenarios  $X$ . As total amount of the continuous rain fall is known to follow an exponential distribution [10], we modeled  $P(X; \Theta_X)$  by an exponential distribution of the total rain fall amount together with a Dirichlet distribution which represents probabilistic allocation of the total amount to the points of the rain fall measurement at 4 dams.

We derived  $P(X|S; \Theta_X, \Theta_{X|S})$  by combining aforementioned  $P(S|X; \Theta_{S|X})$  and  $P(X; \Theta_X)$  through Bayes' theorem, and further developed a probabilistic simulator to compute  $P(X|T_k, S; \Theta_X, \Theta_{X|S})$  from a given  $S$  under the temperature parameter  $T_k$  added for the implementation to the REM.

Figure 4 shows its results of the river floods occurred within a day from the starts of the rain fall. Most of the floods occur within a half day from the starts of the rain fall. In these cases, the floods happen in the midstream which is the intervals 2 and 3 and the downstream which is the interval 5. In contrast, the floods occurred from 20 to 24 hours mainly locate in the downstream. This is because some dams become full at 20-24 hours from the starts of the rain fall and begin to release their water to the downstream. The released water rushes to the downstream together with the other water fallen in the other basins, and cause the floods in the downstream area.

Figure 5 shows the results of the river floods occurred on the 7<sup>th</sup> day from the starts of the rain fall. The times of the flood onsets are diverse, and their locations are also widely distributed from the upstream to the downstream. This diversification is because occurrence of the flood after a large period such as the 7 days from the beginning of the rain fall heavily depends of the rain fall scenario which has many varieties in a long period under complex meteorological conditions. In addition, we observe that the interval 4 of Chikugo river has a sufficient water flow capacity to avoid the flood in the interval.

Furthermore, we evaluated the ratio between probability of flood occurrence and probability of no flood occurrence as  $P(\text{flood onset})/P(\text{no flood})$  for the occurrences on the 1<sup>st</sup> day, the 4<sup>th</sup> day and the 7<sup>th</sup> day respectively. The results are drawn in Fig. 6. As easily understood, the probability of the flood occurrence on the 1<sup>st</sup> day of the rain fall is around one hundredth of the probability of no occurrence of flood in a rain fall occasion. In addition, the probability of a rain fall scenario to cause river flood on the 7<sup>th</sup> day is smaller than that on the 1<sup>st</sup> day in two orders of magnitudes.

These results provide very important insights to plan future investment on the infrastructures for the river control. An important result is that the priority of the investment to the interval 4 should be lower than the others. The main countermeasure required to the floods in Chikugo river is to against the localized torrential downpours rather than the rain fall in a continuous long period.

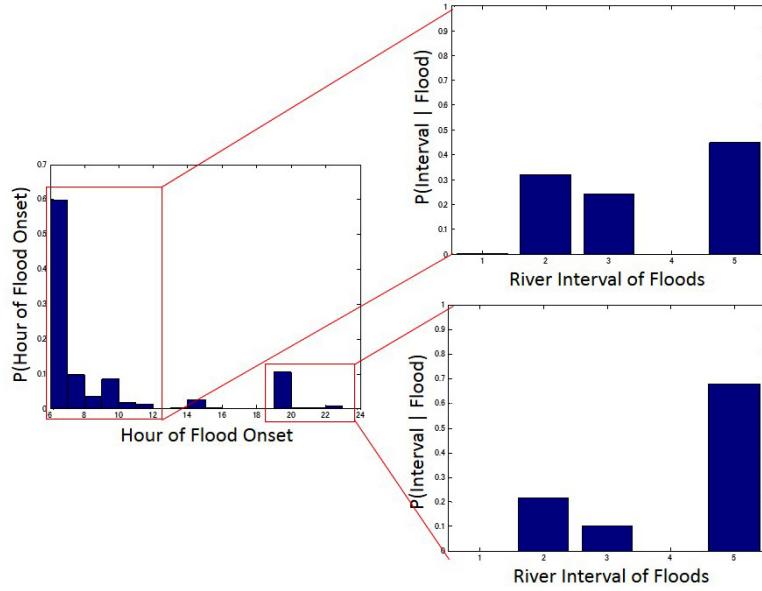


Figure 4 Probability distributions of flood onset times within 24 hours from the rain fall starts (left figure) and probability distributions of flood onset intervals of the river for cases where the onset times are at 6-12hours (upper right figure) and from and at 20-24hours (lower right figure).

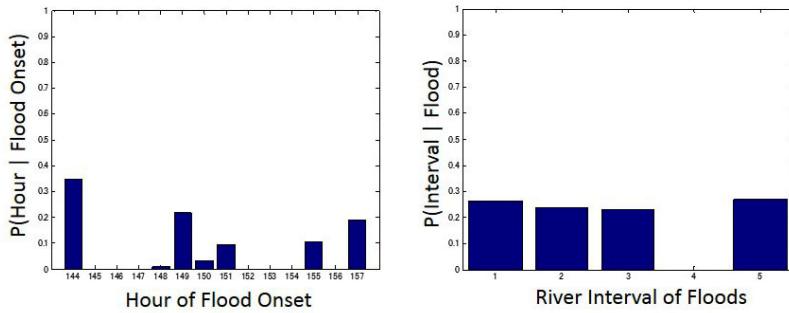


Figure 5 Probability distributions of flood onset times at 144-168 hours (on 7<sup>th</sup> day) from the rain fall starts (left figure) and probability distributions of flood onset intervals of the river for this cases (right figure).

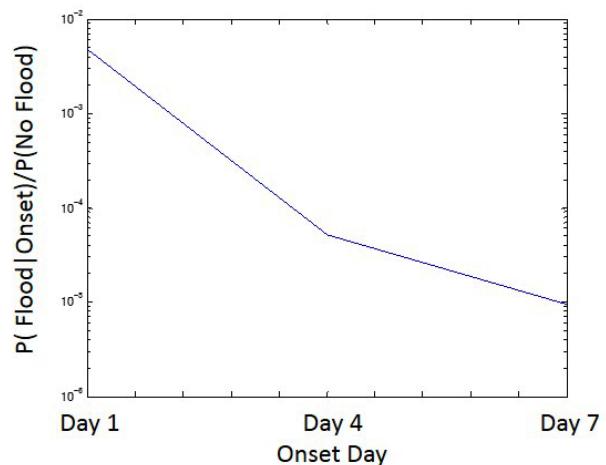


Figure 6 Relative probability distribution of the flood onset days. The vertical axis stands for the ratio between probability of flood occurrences and probability of no flood occurrence.

## (2) Experiments and Results of the Extended Method

In the second year, for the objective (4) and (5), we applied our extended method to estimate extraordinary rain fall scenarios and their probability distribution of the same problem with the first year, which is the occurrence of the severe floods of Chikugo River in Kyushu Island in Japan.

The condition S of a sever flood and its conditioning vector X are defined similarly to these in the first year. The probabilistic simulation model  $P(S|X; \Theta_{S|X})$ , which is the probability of occurrence of a severe flood: S under a given rain fall scenario: X, is identical with that in the first year. Entire basin of the river is partitioned into 8 basins similarly to the first year as shown in Fig.3. The model of each basin is the tank model identical to the first year's. The model of each dam is also defined in the same way with the first year.

Finally, we set up the probability distribution  $P(X)$  by introducing the newly introduced extension. All rain fall data recorded at the 4 dams in the last 10 years do not contain the rain falls causing the severe floods. They are preprocessed in form of the 7 consecutive steps of 4 dimensional vectors to represent the rain fall scenarios X. As total amount of the continuous rain fall is known to follow an exponential distribution [11], we modeled  $P(X; \Theta_X)$  by an exponential distribution of the total rain fall amount together with a Dirichlet distribution which represents probabilistic allocation of the total amount to the points of the rain fall measurement at 4 dams. Subsequently, we calibrated it into  $P(X)$  by using our extension and  $P(S|X; \Theta_{S|X})$ .

We derived  $P(X|S; \Theta_X, \Theta_{X|S})$  by combining  $P(S|X; \Theta_{S|X})$  and  $P(X)$  through Bayes' theorem, and further developed a probabilistic simulator to compute  $P(X|T_k, S; \Theta_X, \Theta_{X|S})$  from a given S under the temperature parameter  $T_k$  added for the implementation to the REM.

Figure 7 shows the entire results of the rare river floods simulation. As shown in the left figure, all floods occur within 140 hours from the starts of the rain fall, and most of them occur within a day from the starts of the rain fall. Majority of the floods appear in the 2nd upstream interval and the 5th upstream (the most downstream) interval as indicated in the right figure. We observe some slight differences between the flood probability distributions provided by the original method in the last year and the method extended in this year.

Figure 8 shows probability distributions of flood onset intervals of the river in the right figure for the cases rounded by a red square in the left figure where the flood onset times are at the 27th hour from the rain fall starts. The frequency of the floods in the 2nd upstream interval is more significant in the result of the extended method than in that of the original method. This is because many floods in the 2nd upstream interval are caused by extremely severe rain falls which probability is very low. Such extremely rare rain falls are underestimated by the original method, while our extended method overcomes this problem as we explained in the former sections. This calibration of  $P(X)$  enhanced the probability of the floods in the 2nd interval.

Figure 9 depicts probability distributions of flood onset intervals of the river in the right figure for the cases rounded by a red square in the left figure where the flood onset times are at the 113th hour from the rain fall starts. The frequency of the floods in the 1st upstream interval is more significant in the result of the extended method than in that of the original method. This is also because many floods in the 1st upstream interval are caused by severe rain falls continued for extraordinary long periods which probability is very low. The calibration of  $P(X)$  in the extended method enhanced the probability in the 1st interval.

These results show that, in general, the extended method provides a larger probability than the original method in case that the events or the scenarios are relatively rarer than the others. This consequence is expected from the calibration scheme  $P(X)=P(X; \Theta_X)/(1- P(S|X; \Theta_{S|X}))$  where X more surely causes S tends to get higher  $P(X)$ . Such X is rarer than the others, if S is a rare target event.

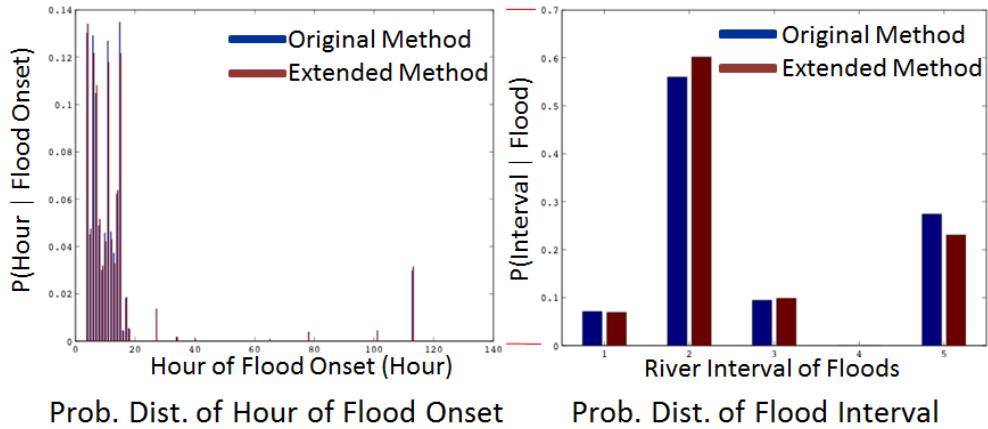


Figure 7 Probability distributions of flood onset times within 140 hours from the rain fall starts (left figure) and probability distributions of flood onset intervals of the river.

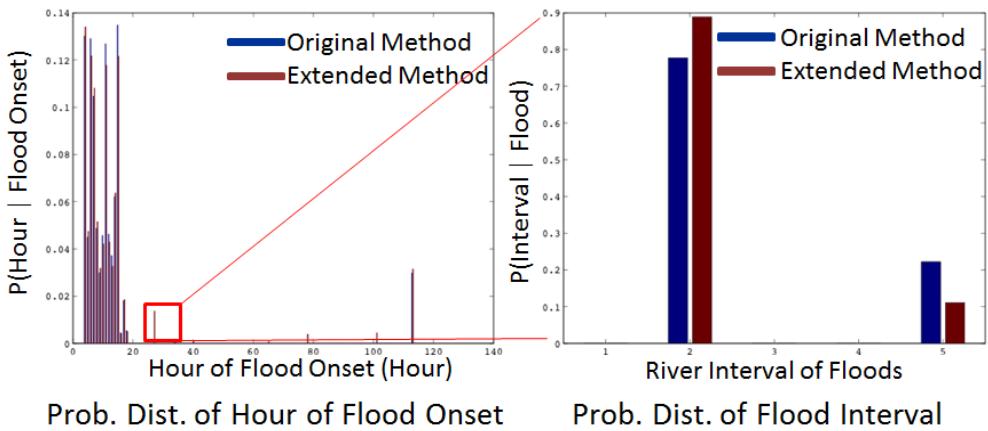


Figure 8 Probability distributions of flood onset intervals of the river (right figure) for the cases which flood onset times are at the 27th hour from the rain fall starts (left figure).

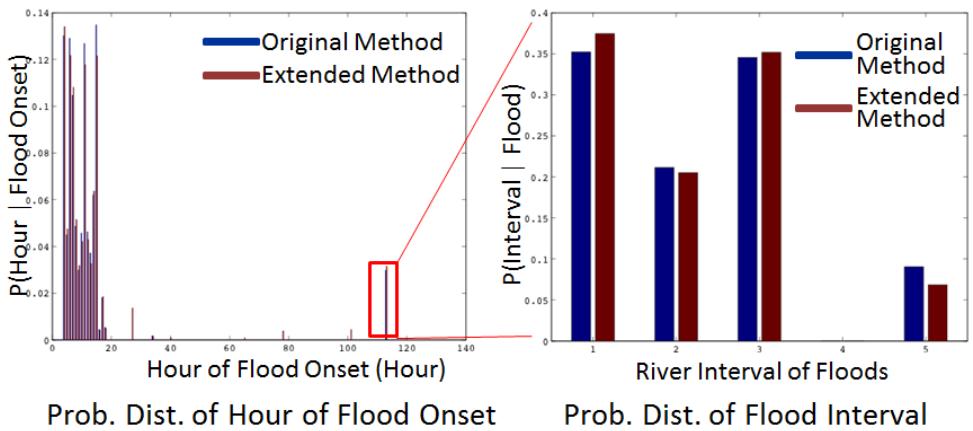


Figure 9 Probability distributions of flood onset intervals of the river (right figure) for the cases which flood onset times are at the 113th hour from the rain fall starts (left figure).

### (3) Experiments and Results of the Integrated Method

In the final year, we applied the integrated method to estimate extraordinary rain fall scenarios and their probability distribution of the same problem with the first and the second years. The condition S of a sever flood and its conditioning vector X, the probabilistic simulation model  $P(S|X; \Theta_{S|X})$ , the model of the entire basin of the river and the model of each dam are defined in the same way with the first and the second years. The integration enables more detailed analysis of the rain fall patterns in both time and space domains.

Figure 10 shows an example rain fall scenario causing the flood in a midstream interval of the river on the 1<sup>st</sup> day from the start of the rain fall. In this scenario, many dams can hold the rain water on the first day, since heavy rain falls are experienced at only one or two dam areas on the day. Accordingly, the effect of the water release of the dams are limited, and the floods almost equally occurs in the both midstream and downstream intervals.

Figure 11 shows an example rain fall scenario causing the flood in a downstream interval of the river on the 1<sup>st</sup> day from the start of the rain fall. In this scenario, heavy rain falls are experienced in more dam areas than the former case where they induce the water release of more dams on the day. Accordingly, the effect of the water release of the dams are more significant. The released water rushes to the downstream together with the other water fallen in the other basins, and cause the more floods in the downstream area. These insights provided by Fig.10 and 11 are consistent with the result of the first year.

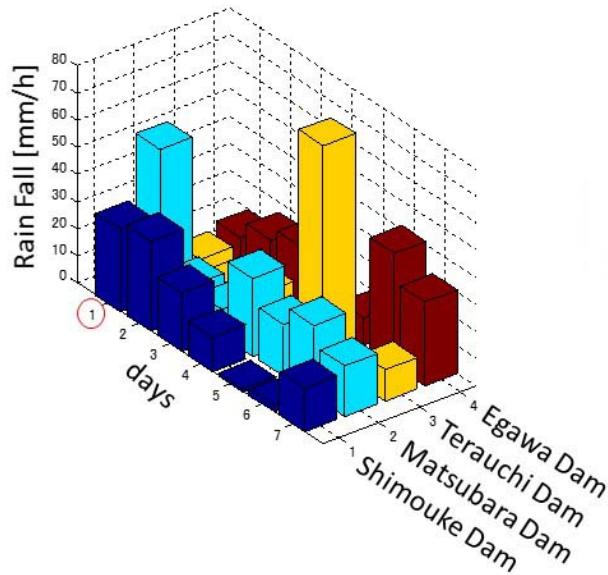


Figure 10 A rain fall scenario at 4 dams for 7 days. This figure shows a scenario to cause the floods in a midstream interval on the 1<sup>st</sup> day from the starts of the rain fall. The vertical axis stands for the rain fall amount per an hour on a day at each dam.

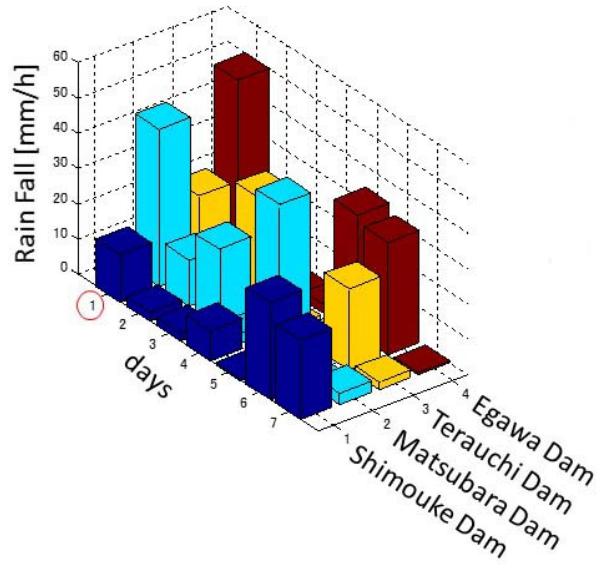


Figure 11 A rain fall scenario at 4 dams for 7 days. This figure shows a scenario to cause the floods in a downstream interval on the 1<sup>st</sup> day from the start of the rain fall. The vertical axis stands for the rain fall amount per an hour on a day at each dam.

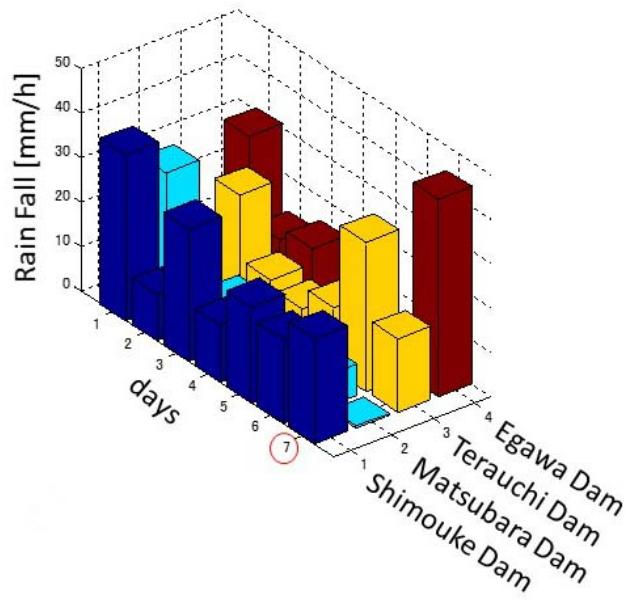


Figure 12 A rain fall scenario at 4 dams for 7 days. This figure shows a scenario to cause the flood in a midstream interval on the 7<sup>th</sup> day from the start of the rain fall. The vertical axis stands for the rain fall amount per an hour on a day at each dam.

A rain fall scenario causing the flood in a midstream interval on the 7<sup>th</sup> day from the start of the rain fall is depicted in Fig.12. Comparing with the former cases, the maximum rain fall amount on a day is rather limited in this case, but the moderately heavy rain falls are experienced in many dam areas for long periods. These rain falls are accumulated in many dams, the dams become full, and finally many dams start to release the water. The released water together with the rain fall water cause the floods after many days from the starts of the rain falls.

**Discussion:**

The results obtained in the first year demonstrated a promising performance of our proposed basic method for risk assessment. It enables detailed and quantitative analyses on the rare and special events/scenarios for both their causes and consequences. Their probabilities are also quantitatively provided based on the mathematically rigorous and probabilistic inference used in our method.

The results of the second year show that the original method can underestimate the probability of the rare events and scenarios by the lack of the past data under the experience of the target event. Moreover, the results indicate the superiority of the method extended in the second year which calibrates this error by our prior knowledge on the rare events/scenarios implemented in the simulation model. This improvement is very important for the risk analysis, since our new technique avoids some fatal underestimation of the probability of severe events/scenarios.

The results of the final year provide more detailed analysis of the events/scenarios causing rare and severe floods of a river in the both space and time domains. The outcomes obtained in the final year give strong impacts to various scientific and engineering domains handling a complex and/or large scale objective system where a complete set of possible events and scenarios in the system is hardly obtained.

A remained issue in this research topic is to further improve the accuracy of the probabilistic simulation models used for both computing the rare events/scenarios and calibrating the underestimation of their probabilities. One measure to address this issue is use of big data which can be acquired from many domains. Even though the big data set does not contain the records of the rare events/scenarios because of the very low frequency of their occurrences, the data is expected to contribute to the improvement of the models from the various aspects.

**References:**

- [1] Rubinstein R.Y. and Kroese D.P.: Simulation and the Monte Carlo Method, 2nd Ed., John Wiley & Sons, Inc., Chap.8, 235-278 (2008).
- [2] Iba, Y. and Hukushima, K.: Testing Error Correcting Codes by Multicanonical Sampling of Rare Events, J. Physical Soc. of Japan, 77, 103801 (2008).
- [3] Hukushima, K. and Iba, Y.: A Monte Carlo Algorithm for Sampling Rare Events: Application to a Search for the Griffiths Singularity, J. Physics: Conf. Series, 95, 012005 (2008).
- [4] Titus S. van Erp: Reaction rate calculation by parallel path swapping, Physical Review Letters, 98(26), 268301 (2007).
- [5] Nguyen V.P., Washio T. and Higuchi T.: New Particle Filter for High Dimensional State Space Models Based on Intensive and Extensive Proposal Distribution, Int. J. Knowledge Eng. and Soft Data Paradigms, 2(4), 284-311 (2010).
- [6] Nguyen V.P. and Washio T.: Modeling Dynamic Substate Chains among Massive States, Intelligent Data Analysis, 12(3), 271-291 (2008).
- [7] Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics), Springer; 1st ed. (2006).
- [8] Binder K. and Heermann D.W.: Monte-Carlo Simulation in Statistical Physics: An Introduction, 2nd ed., Springer (1992).
- [9] Metropolis N. et al.: Equations of State Calculations by Fast Computing Machines, J. Chemical Physics, 21(6), 1087-1092 (1953).
- [10] Shimodaira H.: Improving predictive inference under covariate shift by weighting the log-likelihood function, Journal of Statistical Planning and Inference, 90(2), 227-224 (2000).
- [11] Suzuki A. and Kikuchi H.: A method to estimate a reproduction period of daily rain fall amount under consideration of abnormally heavy rain, J. Weather, 31, 179-189 (1984).

**List of Publications and Significant Collaborations that resulted from your AOARD supported project:**

Papers published in peer-reviewed conference proceedings,

Washio T. and Iba Y., Rare Flood Scenario Analysis Using Observed Rain Fall Data, Proc. JSST 2013; The 32th JSST International Conference on Simulation Technology., Sep. 11-13, 2013, Tokyo, Japan.

Received Outstanding Presentation Award from The 32th JSST Annual Conference International Conference on Simulation Technology (JSST2013).

**Attachments:**

Paper published in peer-reviewed conference proceedings

Washio T. and Iba Y., Rare Flood Scenario Analysis Using Observed Rain Fall Data, Proc. JSST 2013; The 32th JSST International Conference on Simulation Technology., Sep. 11-12, 2013, Tokyo, Japan.